Escuela Politécnica Nacional
Estadística y Probabilidades
Elaborado por: Mat. Fernando Carrasco

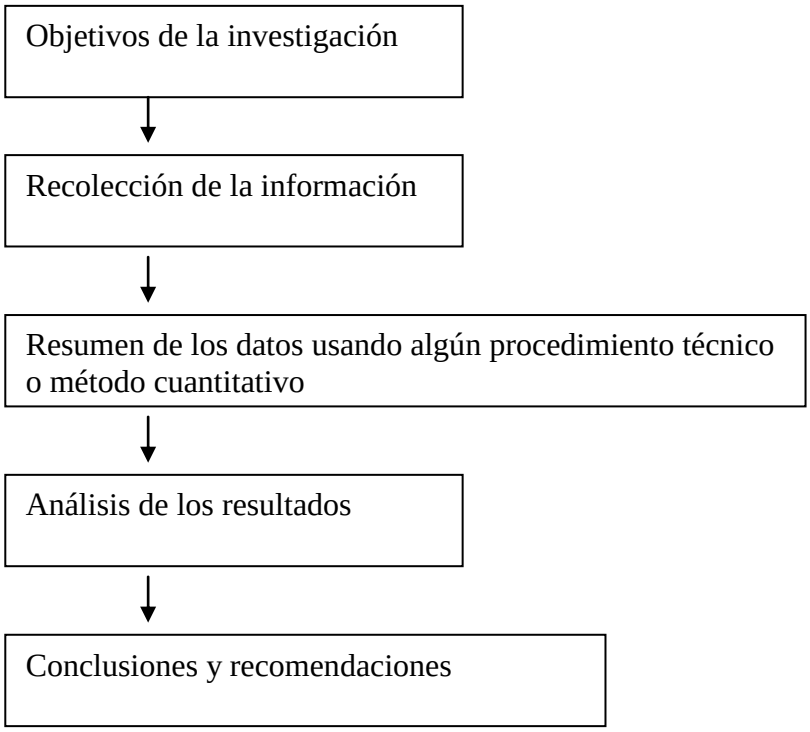
Capítulo 1: Distribuciones de frecuencia

1.1 Introducción.

Al disponer de una cierta cantidad de datos, que han sido tomados de alguna investigación u obtenidos por cualquiera otra forma y, estar interesados en extraer la mayor cantidad de información posible, a partir de la cual podamos realizar algunos análisis y establecer conclusiones, estamos haciendo lo que se llama “ESTADÍSTICA”.

Así, en general podemos decir que la estadística es el proceso de manejo de datos, que comienza por la recolección de la información, el procedimiento de manejo o resumen, el análisis y las conclusiones. Este proceso se lo realiza siempre con base en unos objetivos y por lo tanto, implícito y explícitamente se está realizando investigación de una manera técnica y científica.

Esquemáticamente, tenemos:



Siendo este esquema[1] el proceso general que se sigue en la investigación científica, se tiene que la estadística (que es parte de las matemáticas), sirve para todas las ramas del saber.

[1] Véase, Cumsille Francisco, Investigación, ECO/OPS, Metepec, México 1992.

1.2 Tipos de variables

Al analizar un conjunto de datos, muchas veces no se da ninguna importancia al tipo de datos que se pretende analizar, pudiendo cometer grandes errores al momento de aplicar los procedimientos de análisis, en la interpretación y establecer conclusiones falsas.

Así, tomemos el siguiente caso (muy frecuente de malas interpretaciones y conclusiones).

Supongamos, por ejemplo, que en un cierto día, en la ciudad de Quito, tenemos una temperatura ambiental de 10 grados centígrados (10^oC), mientras que en Guayaquil, hay una temperatura de 20 grados centígrados (20^oC). Con base en estos datos, ¿qué se podría decir o analizar acerca de la temperatura ambiental de las dos ciudades más grandes del Ecuador?.

Algunos dirán:

- en Guayaquil hace más calor (c).
- Quito es más frío (c).
- la temperatura de Guayaquil **es el doble** (20/10=2) que la temperatura de Quito (i).
- la temperatura de Quito **es la mitad** (10/20=0,5) que la temperatura de Guayaquil (i).
- la temperatura de Guayaquil es mayor en 10^oC a la temperatura de Quito (c).
- etc, etc.

Algunas de estas afirmaciones (conclusiones) son correctas (c), pero otras son incorrectas (i) o erróneas.

Pero, **¿porqué son erróneas algunas de dichas respuestas?**. Para verificarlo vamos a dar las mismas temperaturas en otra escala (grados Fahrenheit)

Para la ciudad de Quito:	10^oC	o	50^oF
Para la ciudad de Guayaquil:	20^oC	o	68^oF

Si dividimos 20^oC para 10^oC, aparentemente la temperatura de Guayaquil sería el doble que la temperatura de Quito; sin embargo, si dividimos 68^oF para 50^oF, tendríamos que la temperatura de Guayaquil es 1,4 veces la temperatura de Quito.

Pero la temperatura o la cantidad de calor es la misma independientemente de qué escala se use, luego una persona podría concluir lo siguiente:

- la temperatura de Guayaquil es **dos veces o el doble** que la temperatura de Quito (si tomó los datos en grados C), o
- la temperatura de Guayaquil es **1,4 veces** la temperatura de Quito (si tomó los datos en grados Fahrenheit).

Así, vemos que **estas conclusiones son contradictorias**, y por lo tanto existe un problema con la interpretación de los resultados.

Supongamos ahora que se pesan dos objetos, y se dan sus valores en distintas escalas.

Objeto 1:	1 kg	o	1000 gr	o	2,2 libras
Objeto 2:	2 kg	o	2000 gr	o	4,4 libras

En este caso, sin embargo, si usamos los valores en kilogramos, el objeto 2 pesa el doble que el objeto 1, y de igual manera si usamos los valores en gramos o en libras.

Luego, para este caso, si es válida la conclusión de que el objeto 2 pesa dos veces más que el objeto 1, o que el objeto 1 pesa la mitad que el objeto 2, es decir podemos establecer proporcionalidad entre los valores, mientras que con los valores de temperatura no podemos hacerlo, **¿por qué?.**

La respuesta está en el **tipo o clase de variable** que se está analizando, teniendo que esto juega un rol muy importante en el análisis de los datos, para la correcta aplicación e interpretación de los diferentes modelos, técnicas y procedimientos de la estadística.

Las variables pueden clasificarse así:

- Tipos variables[2]:
  - 1.- Cuantitativas
    - 1.1 Continuas
    - 1.2 Discretas
  - 3.0 Intervalo o escala
  - 2.- Cualitativas :
    - 2.1 Ordinales
    - 2.2 Nominales

Se ha dejado intencionalmente las variables de intervalo o escalas como un grupo aparte por las características y propiedades que tienen estas, abordando primero los otros tipos de variables.
Las variables cuantitativas, son aquellas que se pueden cuantificar o medir, obteniendo valores que cumplen una relación de orden, es decir, que dado dos valores cualesquiera se pueden comparar en el sentido de poder decir que uno de los valores es mayor, menor o igual que el otro valor.

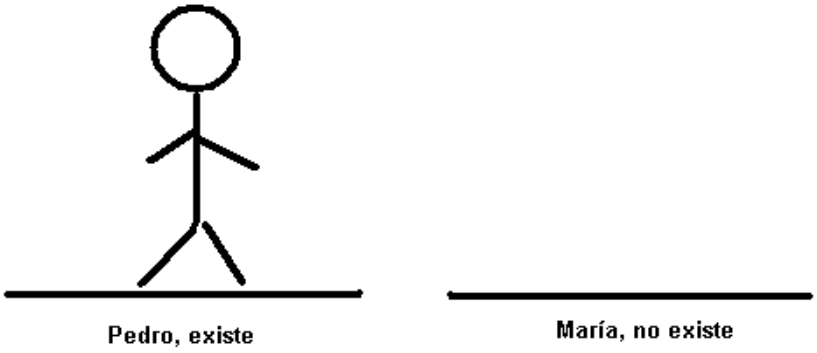
Por otro lado, **las variables cualitativas**, indican solamente una cualidad o característica que no puede ser medida numéricamente.

A las mediciones de la variable cuantitativa se les llama los **valores de la variable**; mientras que, a las observaciones (o mediciones) de una variable cualitativa, se les suele llamar las **categorías de la variable.**

Una de las propiedades más importante de las variables cuantitativas estárelacionado con el valor nulo o cero, el cual representa "**AUSENCIA DE LA CARACTERÍSTICA QUE SE ESTÁ MIDIENDO**", teniendo como consecuencia, la **no existencia del elemento que se estudia.**

[2] Dependiendo de la escuela estadística, esta clasificación puede variar, existiendo en la actualidad cierta polémica y desacuerdos por las propiedades de las variables de escala o interválicas.

Por ejemplo, les presento a “Pedro”, quién mide 1,70 metros, y les presento a María, la que mide 0,0 metros.



Podemos ver a Pedro (existe), mientras que a María no la podemos ver porque simplemente esta persona no existe (ausencia del individuo)

Esta característica, en la que el **valor cero** representa ausencia total del elemento o atributo que se mide, es una **propiedad única y exclusiva de las variables cuantitativas,** y **únicamente en este caso puede establecerse la proporcionalidad** entre valores, es decir, dados dos valores pueden compararse en el sentido de que uno de los valores puede ser el doble, el triple, la mitad, etc., que el otro valor.

Así, retomando la variable temperatura, tenemos que una temperatura de 0^{o}C, no representa ausencia de temperatura (**cero grados centígrados, es un valor de temperatura que lo podemos sentir físicamente**), y por lo tanto no podemos establecer proporcionalidades entre valores de temperatura, siendo este tipo de variables, las denominadas **variables interválicas o de intervalo o de escala**, luego, la variable temperatura no es una variable cuantitativa.

Como ejemplos de variables cuantitativas tenemos, peso, talla, edad, ingresos económicos, etc; mientras que variables como “color”, “nivel socioeconómico familiar”, “profesiones”, “nivel de estudios”, son ejemplos de variables cualitativas.

Dentro del grupo de las **variables cuantitativas**, podemos a su vez diferenciarlas en dos subtipos: las **continuas y las discretas**.

Las variables cuantitativas continuas son aquellas que puede tomar cualquier valor sin ninguna restricción, por ejemplo, podemos tener los valores 5, 6, 5.1, 5.3, 5.12, etc.

Las variables cuantitativas discretas, se caracterizan por tener valores asociados a valores enteros únicamente, por ejemplo, “número de personas en una casa”, pueden haber 1, 2, 3, 4 o 5 personas, pero no pueden haber 2.5 personas.

Por otro lado, a las variables cualitativas, podemos subdividirlas en dos grupos o tipos, **las ordinales y las nominales**.

Las variables cualitativas ordinales se caracterizan por que sus categorías se prestan para un ordenamiento de éstas, mientras que las **cualitativas nominales**, se caracterizan porque no importa el orden en que se presenten sus categorías.

Así, tomando la variable, “Nivel de estudios”, con las categorías

Caso 1 (ordenado)	Caso 2 (no ordenado)
1. Ninguna	1. Secundaria
2. Primaria	2. Postgrado
3. Secundaria	3. Ninguna
4. Superior	4. Primaria
5. Postgrado	5. Superior

Podemos apreciar en ambos casos (1 y 2) que la variable es la misma, puesto que tiene las mismas categorías, sin embargo, parece mejor, tener las **categorías ordenadas (caso 1)**, ya sea en forma ascendente o descendente. Esta característica de ordenamiento y como el mismo nombre lo sugiere, son las variables cualitativas ordinales.

En cambio, variables como: "Color", con las categorías

Caso 1	Caso 2	Caso 3
. Rojo	. Azul	. Verde
. Azul	. Amarillo	. Rojo
. Verde	. Rojo	. Amarillo
. Amarillo	. Verde	. Azul

No tiene ninguna importancia que se registren o presenten las categorías en cualquier orden, siendo éstas las **variables cualitativas nominales**.
Retomando a las variables de escala, donde **el valor cero o nulo no representa ausencia del evento** que se está midiendo, tenemos que esta variable viene a ser como una mezcla de cuantitativa (discreta) con cualitativa (ordinal), la que surge al tratar de cuantificar una característica subjetiva con base en una escala numérica.

Por ejemplo, al querer medir el nivel de factibilidad de un proyecto social, con base en una escala de 0 a 5 puntos (donde 0 representa que no es factible, y el 5 en cambio representa que es totalmente factible), es una variable de escala o interválica.

Así, esta variable la podemos expresar:

"Nivel de factibilidad del proyecto"

0: no es factible
1: poco factible
2: algo factible
3: factible
4: muy factible
5: totalmente factible

Pero aparentemente no se aprecia ninguna diferencia, por ejemplo, con la variable ordinal "Nivel de estudios". Sin embargo, pueden establecerse dos diferencias claras:

- La primera es que con la variable "Nivel de estudios", tenemos estrictamente categorías (cualidades), mientras que la variable "Nivel de factibilidad del proyecto", asigna un valor numérico para tratar de cuantificar la factibilidad de ejecución de un cierto proyecto.

- Con la variable cualitativa ordinal, digamos que el “esfuerzo” o trabajo de ir o pasar de una categoría a otra, es diferente. Por ejemplo, el pasar de la categoría “ninguna instrucción” a la categoría “primaria”, es muy diferente que ir de “superior” a “postgrado”. En cambio, en las variables de escala, este “esfuerzo” es el mismo, es decir, un cambio en un punto es igual al comienzo o al final de la escala, como es el caso de la temperatura.

Con esta última afirmación, sin embargo, hay muchos investigadores que no están de acuerdo, y se están realizando estudios para tratar de llegar a un acuerdo.

Ejemplos de variables de escala son las siguientes: temperatura, coeficiente intelectual, nota de los alumnos en una materia, calificación de un atleta por parte de los jueces, etc.

Luego como el valor cero o nulo, en este caso no representa ausencia del atributo, **con estas variables no podemos establecer proporcionalidad**, pero si calcular algunas estadísticas por tener una escala métrica, como son promedios, varianzas, entre otras.

1.3 Distribuciones de frecuencia

Una vez que se dispone de una cierta de cantidad de datos de alguna variable, interesa analizar esta información, para lo cual el **procedimiento natural consiste en resumir o agregar los datos**, de tal manera que se pueda comprender o extraer la información que estos contienen.

El procedimiento más conocido y fácil, es el llamado **distribución de frecuencias**, que consiste con contar el número de veces que se repite un valor o categoría, obteniendo la llamada “frecuencia absoluta”.

Sin embargo, la manera como se realiza la distribución de frecuencias, depende del tipo de variable que se quiere analizar.

Para el caso de las variables cualitativas, generalmente se procede directamente a contar el número de veces que se repite cada una de las categorías de la variable; pero para el caso de las variables cuantitativas, la mayoría de las veces debe construirse intervalos (llamados a veces “intervalos de clase”).

Así, por ejemplo, tomemos unos datos sobre el tipo de combustible utilizado para cocinar por un grupo de 6167 familias. La tabla de distribución de frecuencias es la siguiente:

Tabla resumen de distribución de frecuencias:
COMBUSTIBLE para COCINAR

		Frecuencia absoluta
Categorías	Gas	4466
	Electricidad	28
	Gasolina	5
	Kérex o diesel	31
	Leña o carbón	1607
	Ninguno /no cocina	30
	Total	6167

Sin embargo, como el objetivo de este procedimiento (distribución de frecuencias), es el de resumir y presentar la mayor cantidad útil y posible de información para los usuarios, a partir de la frecuencia absoluta, suelen construirse otras estadísticas, como:

- La frecuencia relativa, que consiste simplemente en presentar la frecuencia absoluta en términos porcentuales.

- La frecuencia absoluta acumulada, que consiste en ir realizando una suma acumulada de las frecuencias a través de las categorías, ya sea en forma ascendente o descendente. Y, de una forma similar se puede construir también la frecuencia relativa acumulada

Estas últimas estadísticas (frecuencias acumuladas), tienen sentido con todos los tipos de variables descritas, a excepción de las variables cualitativas nominales.

Así, se construyen cuadros resumen de distribuciones que pueden aportar mayor información, como se tiene a continuación:

COMBUSTIBLE para COCINAR

Categorías	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
Gas	4466	72.4%	4466	72.4%
Electricidad	28	0.5%	4494	72.9%
Gasolina	5	0.1%	4499	73.0%
Kérex o diesel	31	0.5%	4530	73.5%
Leña o carbón	1607	26.1%	6137	99.5%
Ninguno /no cocina	30	0.5%	6167	100.0%
Total	6167	100.0%		

Para el caso de variables cuantitativas, tomemos el nivel de ingresos familiares para las mismas familias, en cuyo caso deben construirse intervalos, ya que se tienen demasiados valores distintos, que varían desde 0 hasta 1200 dólares. Puede construirse la siguiente tabla resumen de distribución de frecuencias:

Nivel de ingresos familiares

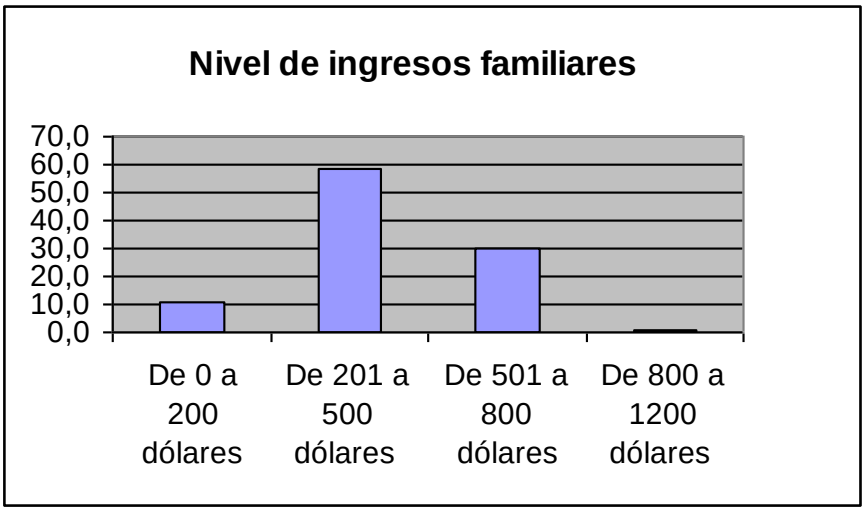
Categorías	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa acumulada
De 0 a 200 dólares	664	10.8	10.8
De 201 a 500 dólares	3603	58.4	69.2
De 501 a 800 dólares	1844	29.9	99.1
De 800 a 1200 dólares	56	0.9	100.0
Total	6167	100.0	

Sin embargo, esta es una distribución totalmente arbitraria, utilizando puntos de corte para la distribución de los datos igualmente arbitraria.

Ante esta situación, ¿existe reglas para establecer el número y la longitud de los intervalos?; la respuesta es que hay una sola regla, y es que "**no hay reglas**". La forma como se construyen estos intervalos depende básicamente de los objetivos de la investigación.

Una cosa que es muy importante resaltar, es que al tener una muestra representativa de la población estudiada, la frecuencia relativa, viene a ser la **probabilidad de ocurrencia**[3] de cada una de las categorías definidas, es decir, se podría decir que la probabilidad "surge" de las distribuciones de frecuencia.

Al realizar una representación gráfica de la distribución de frecuencias, se obtiene un gráfico llamado "histograma de frecuencias":



Este gráfico de barras, nos muestra la distribución de los datos, e implícitamente nos indica cómo se distribuye la probabilidad de ocurrencia de los eventos.

Y, a partir de las formas de los histogramas puede intuirse sobre que distribución o ley de probabilidad se ajusta mejor a los datos para realizar, si fuese el caso, otras pruebas estadísticas.

Hay un caso especial de los histogramas llamado "**Gráfico de PARETO**", el mismo que consiste en graficar las categorías ordenándolas de acuerdo a la frecuencia (de mayor frecuencia a menor frecuencia), conjuntamente con una línea que representa a la frecuencia acumulada. Estos gráficos suelen aplicarse en proyectos de control de calidad, graficando los principales problemas (por su frecuencia) y que tendrán prioridad para su tratamiento.

De estos gráficos de Pareto, surge el famoso criterio llamado "80 - 20", que significa que atacando el 20% de las principales dificultades, puede resolverse el 80% de los problemas.

Nota: para trabajo de clase, desarrollar como se construyen los gráficos de (i) "TALLOS Y HOJAS", (ii) DIAGRAMAS DE CAJA", y (iii) "PIRÁMIDES DE POBLACIÓN".

[3]Algunos investigadores, definen la probabilidad de ocurrencia de un evento, como el límite de la frecuencia relativa, cuando la muestra es suficientemente grande.

Capítulo 2: Medidas de tendencia central

2.1 Introducción.

En el proceso de resumir los datos para describir la información, surgen otros procedimientos englobados en lo que se ha dado por llamar **medidas de tendencia central**, cuyo objetivo o propósito es hallar un valor o categoría que sea "representativo" de todo el conjunto de datos. Este valor "representativo" de las características y atributos de todo el conjunto de datos, es lo que se conoce como el **promedio** de la distribución de datos.

Sin embargo, es muy común observar en libros y publicaciones en general, que dependiendo del área de estudio, por ejemplo, en el campo económico, utilizan la técnica del valor modal para establecer el promedio de los ingresos familiares, o si se revisan aplicaciones en el área de la bioestadística, para establecer el valor promedio de medidas antropométricas, usan el método de la mediana.

Es decir, que el promedio, no consiste, por ejemplo, cuando le preguntan a cualquier persona, ¿cómo calcula la nota promedio de cierta asignatura?, y responde: "se suman todas las notas y se divide para el número de alumnos".

Así, el valor promedio de un conjunto de datos, es un valor que trata de caracterizar o representar a todos los valores, teniendo que al valor promedio se lo suele confundir permanentemente con el procedimiento de la media aritmética.

2.2. Métodos para establecer el promedio de una distribución

El valor promedio, o simplemente el promedio de la distribución de datos, pueden obtenerse por tres métodos o procedimientos:

- la media aritmética
- la mediana, y
- la moda

De estos tres procedimientos, el más conocido y utilizado (por sus propiedades como estadístico) es la media aritmética, y de allí la razón para creer que le media aritmética es el promedio, siendo en realidad que la media aritmética es uno de los procedimientos o métodos para llegar a obtener el valor promedio de nuestro conjunto de datos.

Y, ante la pregunta "¿por qué hay varios métodos para calcular el promedio?", se debe al tipo de variable que se dispone.

Los procedimientos de cálculo del promedio, se definen, considerando que se tiene una variable **X** con una muestra de n valores (X_1,X_2,...,X_n), de la siguiente manera:

- **Media aritmética**: se suman los valores de la variable, y se divide para el número de muestras (o casos) que se dispone, es decir:

$$\overline{X} = \frac{(X_1 + X_2 + X_3 + \ldots + X_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Mediana**: es el valor que divide a la distribución de datos en dos partes iguales. Pero para establecer tal valor, los **datos deben ser primeramente ordenados**, ya sea en forma ascendente o descendente. Así, se tiene que de todo el conjunto de datos, el 50% está por debajo de la mediana, y el otro 50% está por encima de la mediana.

  Así, se ordenan los datos $(X_1,X_2,...,X_n)$

X'_1
X'_2
X'_3 } 50% de los datos
X'_4
X'_5
.
X'_j **este valor es la mediana**
.
.
.
X'_{n-3}
X'_{n-2} } 50% de los datos
X'_{n-1}
X'_n

- **Moda**: se define como el valor o categoría que más veces se repite. Si se llegara a tener dos valores con la misma frecuencia (la más alta), se dice que la moda o el valor modal no existe.

A continuación veamos algunas consideraciones sobre estos procedimientos para establecer el promedio. Se dice que la mediana, "no es sensible a valores extremos", ¿qué significa esto?, veamos un ejemplo: sea la variable edad, con los siguientes valores de años de tres personas:

16
17 (mediana)
18

Calculando la media aritmética, obtenemos 17 años, e igualmente calculando la mediana obtenemos también 17 años.

Supongamos ahora, que en vez de la persona de 18 años, se junta al grupo el famoso personaje bíblico, el señor Matusalén, quién tiene 900 años de edad. Así, la muestra de datos sería la siguiente:

16
17 (mediana)
900

Nuevamente, calculamos la media aritmética, obteniendo una edad promedio para el grupo de 311 años, mientras que la mediana sigue siendo 17 años (porque es el valor que divide al conjunto de datos en dos partes iguales). Luego, podemos apreciar que la mediana no se ve afectada por valores extremos (llamados también, valores fuera de rango), no así la media aritmética, que puede cambiar notablemente, y por supuesto carecer de sentido, siendo en ese caso totalmente inadecuada como una medida estadística de resumen.

Pero, entonces surge la gran pregunta ¿**cuál de los tres procedimientos debe utilizarse, o cuál procedimiento es más adecuado utilizar**?, ante lo cual, lamentablemente, no existe una regla que diga que se debe usar tal o cual método de tendencia central.

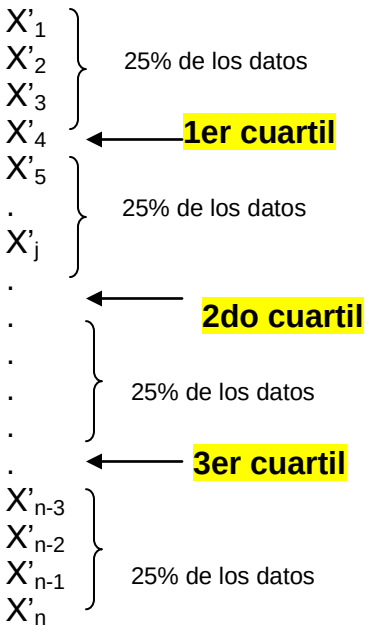
Sin embargo, con base en las distribuciones e histogramas de frecuencias puede tenerse una buena guía para decidir cuál de los tres métodos puede ser el más adecuado, como el caso de la mediana, que al tener una muestra de datos con valores extremos (atípicos o fuera de rango), definitivamente no debe usarse la media aritmética, sino la mediana. Si a pesar de todo, no tenemos la seguridad o confianza de cual medida utilizar, creo que podría tomarse como la mejor guía, la experiencia de miles de investigadores plasmada en libros y publicaciones, donde, de acuerdo al área y tipo de investigación mencionan los procedimientos utilizados.

2.3. Cuantiles

Otras medidas resumen (no de tendencia central), pero sí de posicionamiento a lo largo de la distribución de los datos que ayudan a describir éstos, son los denominados cuantiles, teniendo entre los más frecuentemente utilizados los:

- **Cuartiles**: son los valores del conjunto de datos que dividen a la distribución **ordenada** de datos en **cuatro partes iguales**.
- **Quintiles**: son los valores del conjunto de datos que dividen a la distribución **ordenada** de datos en **cinco partes iguales**.
- **Deciles**: on los valores del conjunto de datos que dividen a la distribución **ordenada** de datos en **diez partes iguales**, y finalmente
- **Percentiles**: son los valores del conjunto de datos que dividen a la distribución **ordenada** de datos en **cien partes iguales**.

Al igual que en el caso de la mediana, los datos primeramente deben ser ordenados, antes de proceder a realizar la partición requerida, por ejemplo, para el caso de los cuartiles, se tendría, ordenando los datos (X_1,X_2,...,X_n):



De forma similar, se procede para los otros tipos de cuantiles. Puede comprobarse fácilmente que el **2do cuartil, es equivalente a la mediana** (puesto que este cuartil divide a los datos en dos partes iguales). De igual forma, el 5to decil o el percentil 50, coinciden también con la mediana, ya que estos valores dividen a la distribución ordenada de datos en dos partes iguales (definición de mediana, medida de tendencia central).

2.3 Indicadores

2.3.1. Introducción

Otras formas de resumir la información, tiene como base los “indicadores”, que tratan de explicar y describir resultados de procesos socioeconómicos en general, permitiendo vigilar de forma periódica el desarrollo e impacto de los procesos de intervención tanto públicos como privados.

Un indicador podría definirse como una variable que busca describir o caracterizar un determinado evento, y su relevancia se lo puede apreciar con base en las siguientes definiciones:

“Los indicadores sociales son estadísticas con un significado y, frecuentemente, con un mensaje. Revelan la realidad detrás de los números. Al develar las diferencias o disparidades en esa realidad, pueden convertirse en herramientas útiles para diagnosticar las desigualdades y seguir el progreso de su erradicación”, Mayra Buvinic, UNESCO

“Los indicadores sociales buscan describir y explicar los resultados del desarrollo social y económico. El desarrollo, como todo proceso de cambio, es producto de la interacción de múltiples factores o causas. Su análisis requiere, por tanto, medidas con distinta capacidad o función explicativa”, SIISE versión 3.5.

2.3.2. Clasificación de los indicadores

Los indicadores pueden ser clasificados en tres clases:

- **Indicadores de resultado**: miden las consecuencias de los procesos socioeconómicos, reflejando los niveles de satisfacción de necesidades básicas alcanzados, pero no reflejan ni explican el porqué del problema.

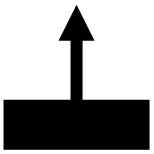
- **Indicadores de insumo**: reflejan los medios necesarios para alcanzar un cierto nivel de vida, es decir la oferta disponible de recursos en general y que son de vital importancia para entender el bienestar, puesto que el hecho de que la población disponga de ciertos condicionantes propios (ingresos, educación, empleo, etc), no necesariamente son una garantía para acceder a los recursos si estos llegaran a ser escasos como los alimentos, sean por causas naturales (terremotos, sequías), o políticas de los gobiernos; o carencia de planteles escolares rurales por falta de inversión pública en áreas de gran productividad.

- **Indicadores de acceso**: miden los determinantes o niveles socioeconómicos propios con que cuenta la población, que condicionan su capacidad para acceder a la oferta disponible de recursos.

  Se recalca que la disponibilidad de medios propios no garantiza el acceso a los recursos (oferta), debido también a costos excesivos de ciertos servicios, como salud privada o educación privada, o acceso a algunos alimentos por restricciones arancelarias.

Esquemáticamente tenemos:

- Indicador de Resultado (consecuencias)
- Indicador de Insumo (recursos disponibles / oferta)
- Indicador de Acceso (condiciones socioeconómicas)



2.3.3. Procedimiento de cálculo: proporciones, razones y tasas

Entre las formas de construir indicadores, tenemos los siguientes procedimientos:

- **Proporciones**: que no es otra cosa que la frecuencia relativa, expresada en términos de la unidad, y por lo tanto puede decirse también que una proporción es simplemente el cuociente de dos cantidades de la misma clase o naturaleza, donde el numerador es una parte del denominador (definición de frecuencia relativa). Cuando a una proporción se le multiplica por cien, se tiene los porcentajes.

- **Razón**: una razón se define como el cuociente de dos cantidades de distinta clase, por ejemplo, el hacinamiento (personas por dormitorio), en el numerador va el número de personas del hogar, y en el denominador el número de dormitorios con que cuenta el hogar.

- **Tasas:** una tasa es una proporción o una razón, multiplicada por una cierta constante para darle más sentido al indicador. Como ejemplo, tenemos el indicador muy conocido “mortalidad infantil”, que se define como el cuociente entre niños menores de 1 año fallecidos y el total de nacidos vivos, multiplicado por 1000, donde la constante nos indica el número de niños muertos por cada 1000 nacidos vivos.

Así, de acuerdo a la definición de tasa, un porcentaje puede ser visto como una tasa. Por ejemplo, es común hablar de “tasa de escolarización de adolescentes” que no es otra cosa que el porcentaje de adolescentes que asisten a clases.

Capítulo 3: Medidas de dispersión

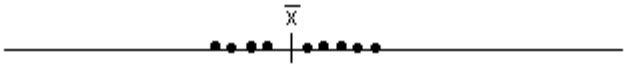
3.1 Introducción.

Supongamos que se tienen dos grupos de estudiantes, los que tienen la misma nota promedio de una cierta asignatura, digamos $\overline{X} = 15$. Ante esto, ¿se podría suponer que el comportamiento de ambos grupos es similar en cuanto al rendimiento académico?.

La respuesta es: “no necesariamente los grupos tienen el mismo comportamiento”, ya que, por ejemplo, podría darse el caso que, en uno de los grupos, todos sus integrantes tengan notas similares y alrededor de la media; mientras que en el segundo grupo, la mitad de ellos son “muy dedicados”, y la otra mitad “muy dejados”, pero que los juntaron en un solo grupo para tratar de motivarlos al estudio.

Gráficamente, esta situación podría verse así:

Primer grupo



Segundo grupo

Puede apreciarse que el primer grupo es homogéneo (calificaciones similares), mientras que el segundo grupo tiene un comportamiento muy heterogéneo (calificaciones muy variables), y por lo tanto tienen comportamientos diferentes estos grupos.

Esto hace necesario analizar la variabilidad de los datos, que viene a ser el último aspecto a considerar en el proceso de describir un conjunto de datos (recuérdese que los dos primeros, digamos, “pasos”, fueron resumir la información con base en las distribuciones de frecuencia, teniendo implícitamente la distribución de probabilidades, y el segundo paso, establecer tendencia central).

Así, con el análisis de la variabilidad de los valores, se viene a cerrar el círculo del análisis descriptivo de un conjunto de datos de manera técnica, disponiendo de elementos de juicio adecuados para el análisis de la información.

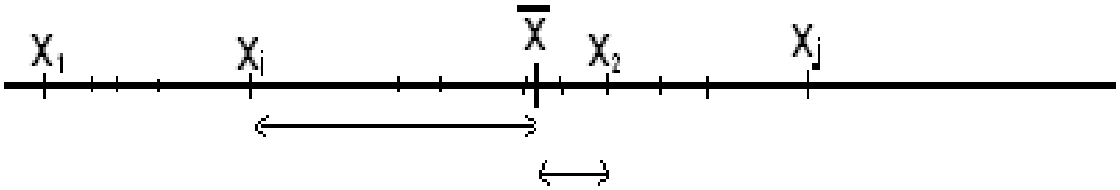
3.2 La varianza y la desviación estándar

El análisis de la variabilidad de los datos juega un rol fundamental en el análisis de los datos, ya que podemos determinar si los datos son relativamente homogéneos o heterogéneos, conceptos de mucha importancia en el tratamiento de datos.

Existen varias estadísticas diseñadas para medir la variabilidad de los datos, sin embargo, vamos a considerar únicamente las dos medidas de variabilidad que se usan cotidianamente, y que además tienen propiedades estadísticas que permiten definir otros conceptos, siendo estas la **varianza y la desviación estándar**.

Para **medir la variabilidad o dispersión de los datos**, se toma un valor de los datos como **punto de referencia**, respecto del cual se cuantifica la variabilidad o dispersión, siendo este punto de referencia, **la media aritmética**.

Así, tenemos el siguiente esquema:



Se define el desvío de un dato cualquiera como la distancia que hay entre la observación y la media aritmética (se toma ésta como punto de referencia), pudiendo tener desvíos positivos (si el dato es mayor que la media aritmética), o desvíos negativos (si el dato es menor que la media aritmética), es decir:

$$desvío = d_i = (X_i - \overline{X})$$

Sea una variable **X** y la muestra de valores $(X_1,X_2,...,X_n)$. A partir de estos desvíos, se definen las medidas de dispersión mencionadas como:

- **Varianza**: (varianza muestral), se define como el promedio de los desvíos elevados al cuadrado.

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- **Desviación estándar**: se define la desviación estándar como la raíz cuadrada de la varianza.

$$S = Desviación\ estándar = \sqrt{Varianza}$$

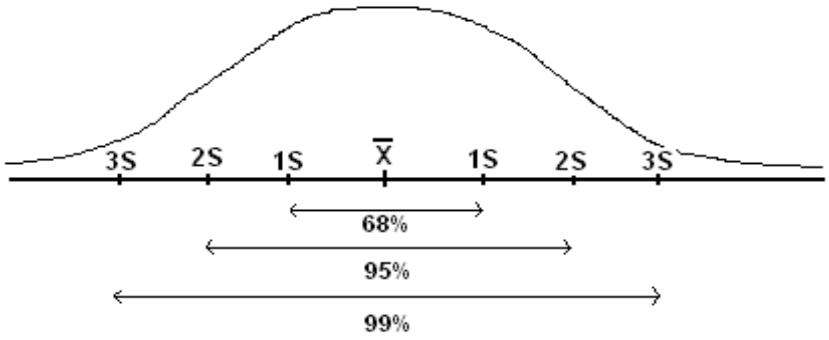
$$S = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Es importante resaltar que la varianza, cuantifica la dispersión de los datos, en la unidad de medida de la variable pero al cuadrado, provocando que sea bastante difícil el poder abstraer o imaginar tal variabilidad, pero no por eso deja de ser útil, ya que ésta tiene un sinnúmero de aplicaciones.

Ante tal situación, como lo deseable es tener una estadística que cuantifique la dispersión de los datos en las mismas unidades que la variable **X**, se obtiene la desviación estándar tomando la raíz cuadrada de la varianza.

La desviación estándar juega un rol vital en el análisis de los datos, ya que, así como para medir longitudes, se tiene el metro como unidad de medida, o para cuantificar pesos, se tiene el kilogramo como la unidad de medida, en el campo de la estadística, interesa cuantificar la dispersión de los datos, teniendo en este caso como **unidad de medida de la dispersión de los datos a la desviación estándar.**

Dada una muestra con un comportamiento en forma de campana, se tiene la siguiente distribución:



Es decir que, en el intervalo que se forma al recorrer una desviación estándar hacia la izquierda y derecha de la media aritmética, se tiene aproximadamente el 68% de todos los datos. Entre dos desviaciones estándar queda el 95% de toda la información, mientras que entre tres desviaciones estándar se tiene el 99% de todos los datos.

Para apreciar la fórmula y cálculo de la varianza, veamos el siguiente ejemplo considerando la variable “Ingresos laborales”:

Datos	Ingresos laborales (X)	(X - media)	elevamos al cuadrado = (X - media)^2	
1	200	(200-324)	15376	
2	250	(250-324)	5476	
3	280	(280-324)	1936	
4	300	(300-324)	576	
5	320	(320-324)	16	
6	350	(350-324)	676	
7	380	(380-324)	3136	
8	380	(380-324)	3136	
9	380	(380-324)	3136	
10	400	(400-324)	5776	
Promedio	324	Sumamos	39240	
		Varianza =	39240 / 9 =	4360

Tomando la raíz cuadrada de 4360, obtenemos la desviación estándar, es decir, 66,0.

Se mencionó que estos conceptos de variabilidad juegan un rol importante en el análisis de los datos, puesto que nos permite establecer el nivel de dispersión y por ende el grado de homogeneidad o heterogeneidad, conceptos de mucha importancia en el tratamiento de datos, puesto que si son homogéneos, entonces el promedio es una medida resumen pertinente y representativa de los datos, en caso contrario no lo es, y por lo tanto no es

adecuado el promedio como una medida de resumen, y en tal caso es preferible presentar una distribución de frecuencias como medida resumen.

Pero como saber si los datos son ¿**homogéneos o heterogéneos**?, la respuesta la tenemos con el llamado coeficiente de variación.

3.3 Coeficiente de variación

El coeficiente de variación (CV) mide la magnitud de la dispersión o variabilidad de los datos respecto del valor promedio, así, este coeficiente se calcula dividiendo la desviación estándar para la media aritmética:

$$Coeficiente\_de\_\text{var}iación = \frac{Desviación\_estándar}{Media\_aritmética}$$

Se acostumbra multiplicar este coeficiente por 100 y, de acuerdo a la experiencia de muchos investigadores y científicos, se ha establecido como criterio de homogeneidad (con base en la experiencia, no por procedimientos técnicos) lo siguiente:

- Si, el coeficiente variación < 20%, se puede asumir homogeneidad
- Si, el coeficiente variación > 20%, se asume heterogeneidad

Otros investigadores, han sugerido como punto de corte, no 20%, sino un 15%, es decir que esto depende del investigador y de los objetivos y necesidades del estudio que se lleve a cabo.

Así, por ejemplo, para el caso de la variable “Ingresos laborales” del punto anterior, tenemos que el coeficiente de variación es:

CV = 66 / 324 = 0,204 (20,4%)

De donde podríamos decir que tal muestra es heterogénea y por lo tanto el promedio (324), **no es una medida resumen adecuada o útil**, puesto que se tiene una población muy variable, y en tal caso como alternativa resumen de los datos es mejor una distribución de frecuencias.

Algo que es importante resaltar de este estadístico (coeficiente de variación), es que **no tiene unidad de medida,**y como tal se lo puede utilizar también para comparar las dispersiones o variabilidades de dos o más variables de naturaleza muy distintas, por ejemplo, el peso y la talla.

Supongamos que quisiéramos conocer de un cierto conjunto de datos, que variable, el peso o la talla, presentan mayor dispersión. Para esto, se calculan los coeficientes de variación de cada una de las variables y se comparan (esto es posible hacerlo, puesto que, el coeficiente de variación, no tiene unidad de medida).

Luego, como el coeficiente de variación mide la magnitud de la dispersión de los datos, comparando los coeficientes, se tiene que el mayor valor, nos indica la variable con mayor dispersión.

3.4 La variable estandarizada

Sea una variable **X** y la muestra de valores $(X_1,X_2,...,X_n)$. Se define la variable estandarizada **Z**, -generada a partir de X-, como:

$$Z = \frac{X_i - \bar{X}}{S} \quad para \quad i = 1,2,3,..,n$$

Propiedades:

i. **La media es cero,** $(\bar{Z} = 0)$
ii. **La varianza es uno,** $(S^2 = 1)$

Este estadístico no tiene unidad de medida y por sus propiedades se la utiliza en varias aplicaciones –algunas se verán más adelante-.

Notas:

1. Se muestra a continuación el uso de este instrumento en la cuantificación del estado nutricional de los niños menores de cinco años para obtener los indicadores llamados "prevalencia de desnutrición crónica", y "prevalencia de desnutrición global". **Esto se desarrolla en clase directamente.**

2. **Se desarrolla en clase también el concepto de covarianza.**